722. The application of wavelets vector quantization of Polish speech

Tarasiuk Mirosław[1], Gosiewski Zdzisław[2]
Department of Automatics and Robotics, Wydział Mechaniczny
Politechnika Białostocka, Wiejska str. 45 C, 15-351 Białystok, Poland
E-mail: [1]*miroslaw.tarasiuk@bialystok.policja.gov.pl*, [2]*gosiewski@pb.edu.pl*

(Received 11 September 2011; accepted 14 February 2012)

Abstract. The paper presents a concept of vector quantization of words uttered in the Polish language. Columns or rows of the matrix obtained as a result of time-frequency analysis of chosen words are vectors used for the further analysis. As a tool in the process of vector quantization was used the Wavelet Packet Transform (WPT) in which the signal decomposition scale is similar to the mel frequency scale (see method - Mel Frequency Cepstral Coefficients – MFCC). Such analysis allowed us to choose the best useful properties for the word recognition. Both column (in time) and row (in frequency) analysis are formulated in the form of computer procedures and compared. We hope such studies will be a starting point for further work on the system Automatic Speech Recognition (ASR).

Keywords: ASR, WPT, PCA.

1. Introduction

The issue concerning ASR is the subject of scientific research conducted by many researchers. Most of the currently used recording equipment change signal processing into a discrete form. In the initial stage of signal processing, it should be separated from the surrounding silence [4]. In the resulting form, {s (n)} it has a large redundancy of information so it is necessary to make the parameterization.

For many years, to parameterize speech signals, the following methods were used: Perceptual Linear Prediction (PLP) and MFCC [15]. They resulted in obtaining a large number of parameters of the input signal. The application of Linear Discriminant Analysis methods (LDA) [7] and Principal Component Analysis (PCA) [6] reduced the number of elements of vector characteristics while limiting the decline in its representativeness [11]. There were also efforts made to segment words [16] in order to continue vectorising a limited number of homogeneous acoustic segments.

Currently, the interest of researchers is focusing on the parameterization based on the use of WTP [2]. Farooq and others [3] proposed the use of decomposition similar to the mel frequency scale. In the range of frequency from 0 to 8 [kHz], 24 subbands were received. In a further signal processing, LDA was used. Kotnik and others [6] used a tree decomposition of the 49 subbands. The use of PCA allowed to limit the features of vectors from 49 to 20. They were assessed during the testing of ASR.

A variety of tree decompositions are tested as well as the methods of assessment of wavelet coefficients obtained. This article presents the results achieved during the parameterization of the speech signal using a new approach to the way of vectorisation, then its parameterization using PCA.

2. Wavelet decomposition

Decomposition of the Wavelet Transformation (called Wavelet Transform – WT) has allowed the use of timescale representation for testing nonstationary signals. In this case, the scale performs an analogous frequency in the Fourier transform (FT) [14]. The major difference

between the two transformations is that FT examines the resolution of the signal with a constant frequency, while for WT frequency, resolution varies with frequency change of the tested signal. This corresponds to the characteristic frequency of human voice route.

A single process signal decomposition means decomposition on the part of the low-pass and high-pass. Subsequent levels of decomposition are obtained by repeating the operation on the part of low-pass. As a result of WT, the number of coefficients is halved (decimation) together with the increased levels of decomposition. If we make a signal decomposition in relation to the low-pass and high pass part, we obtain the WPT.

3. Speech signal parametrisation using WPT

One of the basic methods of parameterization in the framework of human speech recognition is the use of MFCC. In this method, the spectrum obtained by fast Fourier Transform (called Fast Fourier Transform - FFT) is filtered by a filter bank with a width of 300m, shifted relative to each other about 150 mels, covering the entire frequency range. This defined filter bank imitates the human auditory system. In this way, we obtain a vector amplitude or energy signal vector, which, after logarithmising, we subject to discrete cosine transform and obtain a vector of central factors that contain as many parameters as there are mel bands.

Sarikaya and others [13] in the system recognising the speakers have used a 6-Level WPT, where, gradually moving to the lower levels of tree decomposition, they obtain the frequencies corresponding to the filter bank of MFCC method. At the same time, they reduce the number of considered frequency subbands from 64 to 24. The diagram of tree decomposition for the frequency sampling 8 [kHz] is shown on Fig.1. This figure is simplified by omitting the names of subbands and frequency ranges. The colour highlights the used decomposition.

Since with this method of decomposition, we obtain 4 kinds of frequency bands with different amounts of wavelet coefficients, action should be taken to allow for their comparison. We propose the designation of an averaged energy of each of the bands according to the formula (1) and use the decomposition method from Fig. 1 for speech recognition:

$$D[k] = \frac{1}{N_k} \sum_{n=1}^{N_k} d_n^2, \text{ where } k = 1,...24 \tag{1}$$

where: N_k is the number of wavelet coefficients in k-filter, d_n - wavelet coefficient in the band.

The speech signal can be divided into short fragments of 10 : 40 [ms], assuming their quasi-stationarity [17]. This follows from the fact that human ear does not respond to changes in a short acoustic signal. These fragments can then be replaced by the parameters calculated from the received frame of observation. The use of frames causes the distortion of actual signal parameters. This is due to the creation of processed signal discontinuities and the emergence of additional high-frequency in its spectrum. To avoid these difficulties, Hamming window should be used to absorb the side samples of signal. It is appropriate to use the windows 30 [ms] with 10 [ms] frames overlap (overlapping).

Using WPT with the proposed earlier (Fig. 1) way of decomposition and averaged energy spectrum (model 1), we obtain for each window a set of 24 energy values in each frequency band. After applying this operation for the whole word, we get the matrix whose rows correspond to the number of windows that were used to separate words. At the same time, the 24 columns correspond to the bands of frequencies in the range from 0 to 4 [kHz]. We carry out the parameterization process for a particular word ‘two’, that has been separated from the surrounding silence.

The word is divided into frames of observation, which employ WPT using 'dB6' wavelet. After averaging the energy in the bands 14x24, we obtain the matrix which is shown in Figure 2.

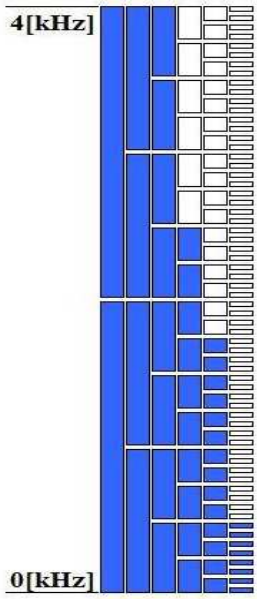

Fig. 1. Six - level package signal decomposition with highlited distribution of the frequency bands similar to MFCC

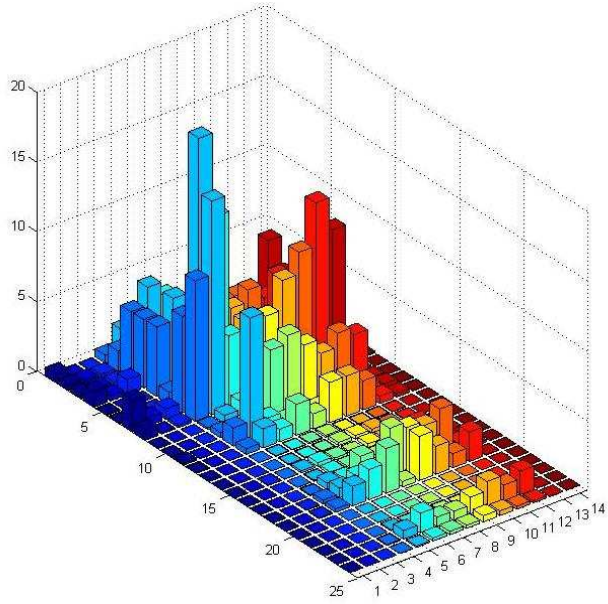

Fig. 2. Graph of parameterized word 'two'

4. Principal components analysis

To reduce the dimensionality of data, we can use PCA [1, 10]. We have input data in the form of a matrix of real numbers, where the rows (vectors with features x_1, ..., x_p), are considered as vectors of observation. A sequence of n vectors represents the tested word:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{2}$$

For the observation vector, we are looking for such a linear combination,

$$z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = a_1^{'}x \tag{3}$$

which is the maximum sample variance.

$$s_{z1}^2 = a_1^{'}Sa_1 \tag{4}$$

S is the sample covariance matrix of x_1, x_2, ..., x_p, additionally vector a_1 satisfies the normalization condition $a_1^{'}a_1 = 1$, ensuring uniqueness of the main component. Vector a_1 is a vector characteristic of the corresponding largest value λ_1 of the matrix S, otherwise the largest root of the characteristic equation:

$$| S - \lambda I | = 0 \tag{5}$$

In doing so, we obtain the successive observations:

$$\begin{aligned} z_1 &= a_1^{'} x \\ z_2 &= a_2^{'} x \\ &\ldots\ldots\ldots\ldots \\ z_n &= a_n^{'} x \end{aligned} \tag{6}$$

Determination of the covariance matrix starts with determining the value of the average *i*-features [12] calculated using the formula:

$$\bar{x}_i = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \text{ , where } j = 1, 2, \ldots, p \tag{7}$$

To determine the characteristics of the variance, can be used unbiased estimator:

$$s_j^{\,2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{ij})^2 \text{ , where } j = 1, 2, \ldots, p \tag{8}$$

For a very large number of samples, both values of the estimators are practically equivalent [8], however, with an unknown distribution of data, the unbiased estimator [5] should be used. To determine the coefficients of covariance [12], the following formula is used:

$$S_{jk} = \frac{1}{n-1} \sum_{\substack{j=1 \\ k=1}}^{p} (x_{ij} - \bar{x}_{ij})(x_{ik} - \bar{x}_{ik}) \tag{9}$$

Since the covariance matrix is a symmetric matrix (S_{12}=S_{21}), it is enough to confine to determining the coefficients of the diagonal and the upper half of the matrix, and then the missing values are filled:

$$S = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1p} \\ \ldots & S_{22} & \ldots & S_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & S_{pp} \end{bmatrix} \tag{10}$$

The next step is to determine the values of covariance matrix by solving the equation (5). Once defined, the value of their eigenvectors is determined, which must satisfy the equation:

$$CV = VD \tag{11}$$

where: V - matrix eigenvectors of the covariance matrix, D - diagonal matrix of eigenvalues.

At this stage, we evaluate the number of main components with which input data are imaged [9]. It can be done by using the percentage of variation of the input vector *X* restricted to the first *k* principal components in relation to the whole:

$$\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p} 100\% \tag{12}$$

Then we make the projection of input data on *k* eigenvectors (sorted in descending order). Thus, we go into space with a reduced number of the features of vectors, the new coordinate system:

$$y = (x - \bar{x}) V_k^T \tag{13}$$

where: $(x - \bar{x})$ – centered input data matrix, V_k^T – transposed matrix of *k* eigenvectors.

5. PCA use for reducing characteristics of vectors describing speech signals

As a result of WPT, we get a matrix of parameters of the speech signal. Logically, two ways of secretion vectors can be considered.

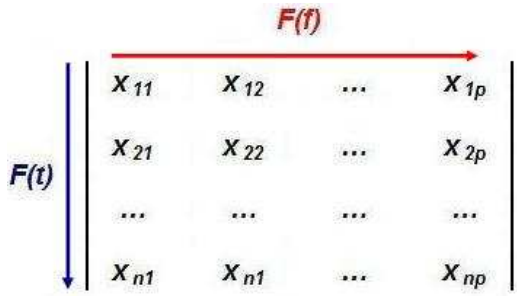

Fig. 3. Two methods for tracing the matrix of parameters

If we get the vector at a given moment of time, then it has *p* describing qualities that determine the energy level in the frequency band (rows). Thus, we obtain *n* vectors. We can then define that it is built on frequency function. Another way of describing it is the *p* vectors occurring in this frequency range (columns). The vectors are defined by energy levels occurring at a given moment of time (*n* features) from the beginning to the end of the word. Then we describe the vector using the function of time. Both ways of vectorization have been checked to reduce the number of features using PCA.

The data obtained from the parameterization of the word 'two' (Fig. 2), contain 14 vectors described by 24 features (frequency). The Covariance matrix has the size of 24x24. After determining the eigenvalues of the covariance matrix, an assessment of their number is made (Table 1a).

The use of the two main components of the variables will represent 84.82% of variance in the variance of input data. If we use the 5 main components of the variables, it will use 96.94% of the variation. In this case, we obtain the resulting matrix in the size of 14x5. Registered input signal after conversion in a system with five main components will look like in Fig. 4a.

If we consider that the vectors describing the word appear in the columns, we obtain 24 vectors. Each vector will describe the emergence of energy in the frequency band in subsequent

moments of time (a function of time). In the case of a word from Fig. 2, each vector will be represented by 14 features. The Covariance matrix will have dimensions of 14x14. After determining the eigenvalues of the covariance matrix, an assessment of their number is made (Table 1b).

With this definition of the vector, using the 2 main components of the variables will represent 89.34% of the variation of input data. The use of 5 main components of the variables results in the use of 97.88% of the variation. Registered input signal after conversion in a system with five main components will look like in Fig. 4b. When you move to a new coordinate system, the word will be described by 24x5 matrix.

Table 1. Evaluation of their initial 12 values

a) Function of 'frequency'				b) Function of 'time'			
No	Eigenvalues	Cumulative value	Cumulated %	No	Eigenvalues	Cumulative value	Cumulated %
1	54,63	54,63	69,53	1	46,84	46,84	73,13
2	12,02	66,65	84,82	2	10,38	57,22	89,34
3	5,27	71,93	91,53	3	2,65	59,87	93,48
4	2,63	74,55	94,87	4	1,93	61,81	96,50
5	1,62	76,17	96,94	5	0,89	62,69	97,88
6	0,91	77,08	98,10	6	0,51	63,21	98,69
7	0,90	77,98	99,24	7	0,49	63,69	99,45
8	0,39	78,37	99,74	8	0,22	63,91	99,79
9	0,10	78,47	99,87	9	0,07	63,98	99,89
10	0,05	78,53	99,93	10	0,03	64,01	99,94
11	0,04	78,56	99,98	11	0,02	64,03	99,98
12	0,02	78,58	100,00	12	0,01	64,05	100,00

a) b)

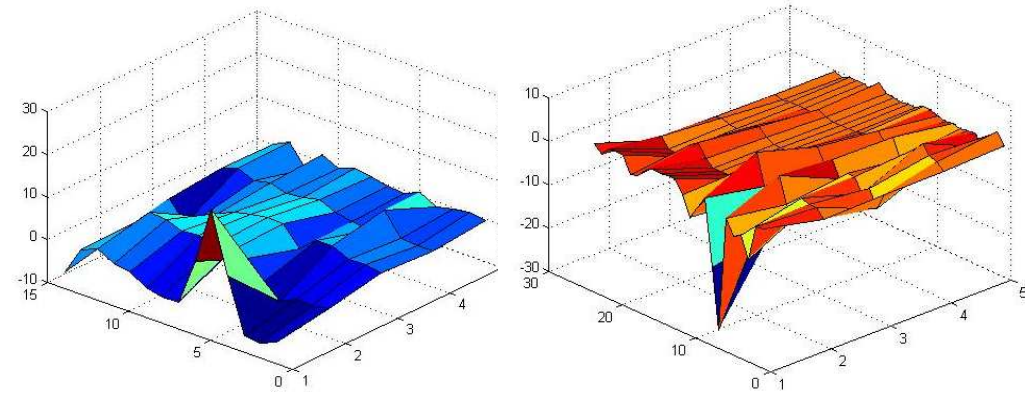

Fig. 4. The word 'two' after the transformation: a) based on five major components of the frequency, b) based on five major components of the time

6. Research

It is necessary to check which ways of the description of vector qualities after their reduction using PCA retain their representation at higher levels.

The study involved two research groups of 10 men and women of different ages, expressing the numbers from 1 to 10. All the material was recorded on the same recorder (Sony ICD-P28 model) with a sampling rate of 8 kHz.

Each recorded word, after using the parameterization described above, was twice subjected

to PCA. In the first case, it was assumed that the vector had 24 features described by frequency bands. In the second case, 24 vectors were investigated. They were having the same number of features as the equivalent of windows a registered word was divided into. It was checked which method of words vectorisation - after time or frequency, used more variability of the data input in the transition to the new coordinate system. It is also necessary to determine what quantity of the variables of principal components is essential for the proper projection of input data.

The percentage results of the cumulative value of eigenvalues set in the two study groups are shown in Fig. 5. Each word was assigned eigenvalues in both cases. Then the minimum, maximum and average eigenvalue values were established, limiting in the figures to the first 5 values sorted in descending order.

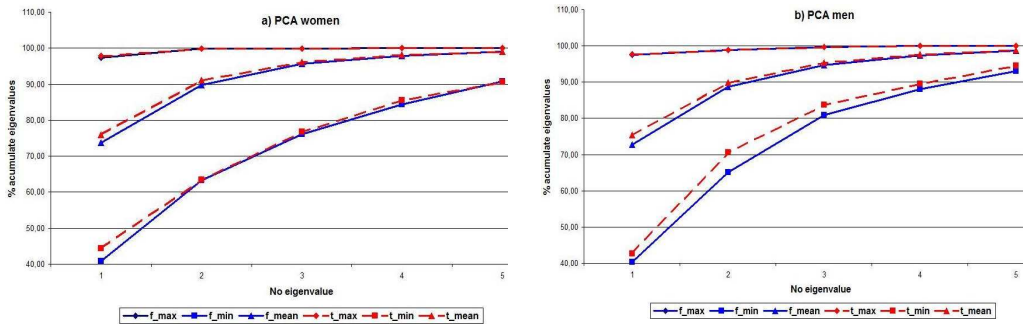

Fig. 5. a) The eigenvalue of a group of women, b) the eigenvalue of a group of men

Coloured highlights: a broken, red - the results obtained by the frequency vector, and blue - by the time. Using the first 5 eigenvalues provides a medium for women - 98.97% of variance in the variance of the input variables in after frequency tracing. In the case of after time tracing, we get 99.08%. Using the first 5 eigenvalues provides a medium for men - 98.73% of variance in the variance of the input variables in after frequency tracing. In the case of after time tracing, we get 99.90%. It is advisable to use five eigenvalues, which will ensure 90% representation of the variance of input data in a transition to the new coordinate system.

7. Summary

The research has shown that the applied algorithm parameterization using the proposed distribution of WTP can be used for the speech analysis. Obtained with its help the matrix representing the tested words can be vectorised after 'time' and 'frequency'. In both cases, the reduction features vectors using PCA in both study groups.

The comparable results have been reached with a slight superiority of the after 'time' method, particularly visible when considering the minimum and medium results.

In the case of after 'frequency' tracing, we obtain the reduction of the number of frequency bands from 24 to 5, but every word is represented by a variable number of vectors. In the case of after 'time' tracing, we get 24 vectors with five characteristics. Each word is then represented by a matrix of the same dimensions.

In further research work, it should be verified which vector method is simpler to implement or achieves a higher level of correctly recognized words.

Acknowledgements

Scientific work co-financed by European Social Fund and from the State Budget and the Budget Funds of Podlaskie Province under the Project 'Innovation Strategy in Podlaskie Province - Building Implementation System'.

References

[1] **Balicki A.** Systematyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne. Wydawnictwo Uniwersytetu Gdańskiego**,** 2009, (in Polish).
[2] **Daubechies I.** Ten Lectures on Wavelets. SIAM, USA, 1997.
[3] **Farooq O., Datta S.** Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Signal Processing Letters, Vol. 8, No. 7, 2001.
[4] **Gosiewski Z., Tarasiuk M.** Preliminary study of the automatic speech recognition for devices supporting the ill and disabled. Journal of Vibroengineering, Vol. 11, No. 3, 2009, p. 497-503.
[5] **Jóźwiak J., Podgórski J.** Statystyka od Podstaw. Polskie Wydawnictwo Ekonomiczne, Wyd. VI, 2006, (in Polish).
[6] **Kotnik B., Kacic Z., Horvat B.** The usage of wavelet packet transformation in automatic noisy speech recognition systems. IEEE Eurocon, Slovenia, 2003.
[7] **Kotnik B., Kacic Z.** A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and AR modeling of speech signals. Signal Processing, Vol. 87, 2007, p.1202-1223.
[8] **Kotulski Z., Szczepiński W.** Rachunek Błędów dla Inżynierów. WNT, 2004, (in Polish).
[9] **Koronacki J., Ćwik J.** Statystyczne Systemy Uczące Się. WNT, 2005, (in Polish).
[10] **Krzyśko M., Wołyński W., Górecki T., Skorzybut M.** Systemy Uczące Się. Warszawa, WNT, 2008, (in Polish).
[11] **Kwon O. W., Chan K., Lee T. W.** Speech feature analysis using variational Bayesian PCA. IEEE Signal Processing Letters, Vol. 10, No. 5, 2003, p.137-140.
[12] **Oktaba W.** Metody Statystyki Matematycznej w Doświadczalnictwie. PWN, 1986, (in Polish).
[13] **Sarikaya R., Pellom L. B., Hansen J. H. L.** Wavelet packet transform features with application to speaker identification. NORSIG'98, 1998, p. 81-84.
[14] **Siafarikas M., Mporas I., Ganchev T., Fakotakis N.** Speech recognition using wavelet packet features. Journal of Wavelet Theory and Applications, Number 1, 2008, p. 41-59.
[15] **Shrawankar U., Thakare V.** Feature extraction for a speech recognition system. In Noisy Environment: a Study. Second International Conference on Computer Engineering and Applications, ICCEA, 2010, p. 358-361.
[16] **Tarasiuk M., Gosiewski Z.** Segmentacja mowy polskiej z wykorzystaniem transformacji falkowej. Acta Mechanica et Automatica, Vol. 4, No. 1, 2010, p. 92-95, (in Polish).
[17] **Zieliński T. P.** Od Teorii do Cyfrowego Przetwarzania Sygnałów. Zakład Poligraficzny Uniwersytetu Jagielońskiego, 2002, (in Polish).